Estimating the Information Gap between Textual and Visual Representations

Henning, Christian Andreas ; Ewerth, Ralph

Abstract: Photos, drawings, figures, etc. supplement textual information in various kinds of media, for example, in web news or scientific publications. In this respect, the intended effect of an image can be quite different, e.g., providing additional information, focusing on certain details of surrounding text, or simply being a general illustration of a topic. As a consequence, the semantic correlation between information of different modalities can vary noticeably, too. Moreover, cross-modal interrelations are often hard to describe in a precise way. The variety of possible interrelations of textual and graphical information and the question, how they can be described and automatically estimated have not been addressed yet by previous work. In this paper, we present several contributions to close this gap. First, we introduce two measures to describe cross-modal interrelations: cross-modal mutual information (CMI) and semantic correlation (SC). Second, a novel approach relying on deep learning is suggested to estimate CMI and SC of textual and visual information. Third, three diverse datasets are leveraged to learn an appropriate deep neural network model for the demanding task. The system has been evaluated on a challenging test set and the experimental results demonstrate the feasibility of the approach.

DOI: https://doi.org/10.1145/3078971.3078991

Estimating the Information Gap between Textual and Visual Representations

Christian Henning[◇]
[◇]Leibniz Universität Hannover
Institute of Distributed Systems, and
L3S Research Center
Hannover, Germany
christian.henning@tib.eu

Ralph Ewerth[◇,★]
[★]German National Library of Science & Technology (TIB)
Department of Research and Development
Research Group Visual Analytics
Hannover, Germany
ralph.ewerth@tib.eu

ABSTRACT

Photos, drawings, figures, etc. supplement textual information in various kinds of media, for example, in web news or scientific publications. In this respect, the intended effect of an image can be quite different, e.g., providing additional information, focusing on certain details of surrounding text, or simply being a general illustration of a topic. As a consequence, the semantic correlation between information of different modalities can vary noticeably, too. Moreover, cross-modal interrelations are often hard to describe in a precise way. The variety of possible interrelations of textual and graphical information and the question, how they can be described and automatically estimated have not been addressed yet by previous work. In this paper, we present several contributions to close this gap. First, we introduce two measures to describe cross-modal interrelations: cross-modal mutual information (CMI) and semantic correlation (SC). Second, a novel approach relying on deep learning is suggested to estimate CMI and SC of textual and visual information. Third, three diverse datasets are leveraged to learn an appropriate deep neural network model for the demanding task. The system has been evaluated on a challenging test set and the experimental results demonstrate the feasibility of the approach.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; *Natural language processing*; *Computer vision*; Information extraction; Natural language generation; Computer vision tasks; Computer vision representations;

KEYWORDS

Text-image relations; multimodal embeddings; deep learning.

ACM Reference format:
Christian Henning[◇] and Ralph Ewerth[◇,★]. 2017. Estimating the Information Gap between Textual and Visual Representations. In *Proceedings of ICMR '17, June 6–9, 2017, Bucharest, Romania, ,* 9 pages.
DOI: http://dx.doi.org/10.1145/3078971.3078991

Christian Henning[◇] and Ralph Ewerth[◇,★]

ICMR '17, June 6–9, 2017, Bucharest, Romania
 ACM ISBN 978-1-4503-4701-3/17/06.
DOI: http://dx.doi.org/10.1145/3078971.3078991

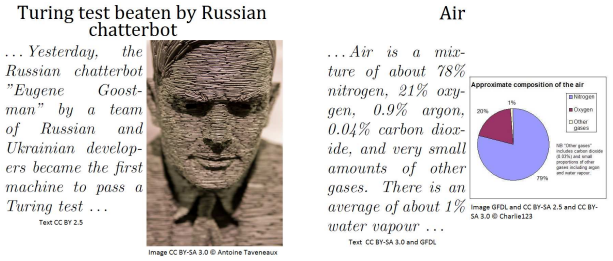
1 INTRODUCTION

"A picture is worth a thousand words." This insight is often utilized to enhance, for instance, textual information in documents, where photos, graphics, diagrams are inserted to supplement textual information. But a bunch of other cross-modal interrelations exists. For example, supplementing a video sequence with music, overlaid speech, and/or overlaid text is very common. In the early stage of (silent) movies, text inserts were used to complement visual scene content with additional text information. Interestingly, in the very early stage of film, text inserts (intertitles) were even used to inform the audience about what will happen in the subsequent shot. This, however, changed soon and intertitles were used in a much more creative and complementary way[1]. Talks in lectures or scientific presentations are complemented with slides[2] which themselves often consist of a mixture of textual, visual, and audio-visual information. Also in the field of software engineering visual representations are exploited, e.g., via specialized diagrams based on the Unified Modeling Language, to describe and understand complex software architectures. These examples hint at the power – but also at the complexity – of combining two or more modalities to convey information in a more appropriate or more understandable way. In this respect, this paper focuses on describing and measuring interrelations of textual and visual information, in short, image-text relations.

In general, two or more different modalities can be used to convey information in a better way. On the other hand, an additional modality (or communication channel) does not always provide an improvement by means of information gain. It can be observed that complementary information is often added for aesthetic reasons or as a visual anchor, e.g., in Web news. For example, textual and visual information is often more related to one another in scientific documents than it is in Web news (cp. Figure 1), we will get back to that later. But how can we describe precisely which and how much information is shared by a text and a related image? How can we describe if the visual information emphasizes an aspect of the text or vice versa, and how can we measure by which means textual and visual information are used complementary?

In recent years, Natural Language Processing (NLP) and Computer Vision (CV) have been employed and combined to tackle interesting challenges that are somewhat related to these questions, for instance, automatic image captioning or multimodal document retrieval. Increasing computational power and deep learning have

[1]In fact, during the first Academy Awards ceremony in 1929, an Oscar was awarded for *Best Writing – Title Cards*, but there was never again an award for intertitles.
[2]Though their usefulness might be questionable in some cases.



(a) News articles are often loosely tied to their co-occurring image.

(b) Encyclopedia articles are typically mainly concerned about the content of an enclosed image.

Figure 1: Examples of image-text pairs with low (a)[3] and high (b)[4] semantic correlation.

enabled impressive systems such as NeuralTalk2 [12] that generates image descriptions in real-time. However, the focus of these approaches is to generate *precise descriptions* of the scene content that is depicted in an image, whereas they do not aim at providing complementary information. Some approaches investigate the interrelations of text and images, but they simply assume that an image is always semantically related to its surrounding text [23, 27]. For the latter, it is easy to find examples showing that this is clearly not true, for example, by simply exploring today's online news in the World Wide Web. On the other hand, automatic recognition of image-text interrelations and correlations in terms of quantity (how much information is shared) and quality (how much meaning is shared) would open up new possibilities to address a variety of interesting challenges and applications: For example, news retrieval can be optimized by selecting articles that include images that show *specific* aspects for a topic or providing a diverse result set, respectively.

In this paper, we present two main contributions that aim at illuminating the gap between – as well as relations of – textual and visual information[5]. First, two measures are introduced that describe different aspects of image-text interrelations: Cross-modal Mutual Information (CMI) and Semantic Correlation (SC). *Cross-modal Mutual Information* captures the amount of information that is shared, while *Semantic Correlation* measures how much meaning is shared among text and image. In Section 3, it is explained and derived why two complementary measures are required to describe image-text relations. As a second main contribution, we present a novel deep learning framework to automatically estimate image-text interrelations by means of CMI and SC. This deep learning framework consists of an autoencoder that exploits a multimodal embedding in order to gather a compact representation of a multimodal document, i.e., a text (document) plus an image in our case. The network uses the popular *InceptionV3* [19] model to encode images. The overall encoding architecture is based on the *Neural Image Caption Generator* from Google [20], which we have extended by an hierarchical text encoding that enables the comprehension of sentences as entities appearing in the context of a variable-sized text. The compact multimodal embedding representation is finally used to train a classifier in order to infer CMI as well as SC. Experimental results demonstrate the feasibility of the approach on a challenging test set including Web news and Wikipedia pages.

The remainder of the paper is structured as follows. Related work is presented in Section 2. Section 3 motivates, introduces and explains in detail the two measures CMI and SC for describing image-text relations. Subsequently, the deep learning architecture for automatically estimating image-text relations is explained in Section 4. The experimental results are presented in Section 5, while Section 6 concludes the paper and outlines areas for future work.

2 RELATED WORK

Many researchers have moved their focus onto tasks involving multiple modalities, especially for tasks laying on the edge of Natural Language Processing and Computer Vision. One interesting idea is to think of the syntax and semantics of images and texts, respectively, being arranged in a hidden latent space, where both modalities can be projected to a multimodal embedding space [7, 8, 11–14, 20, 24].

The approach which proved to be most promising involves deep neural networks to generate the embedding space. A more fine-grained approach has been proposed by Karpathy and Li [12] and Karpathy et al. [11]. They intentionally assimilate decompositions of their representations (in addition to the full input) and ensure that those match up in the embedding space as well.

Vinyals et al. [20] use a simpler approach, motivated by recent advances in statistical machine translation [21]. They generate image captions by transforming an image to a compact representation (a fixed embedding) via deep CNNs (convolutional neural networks) and then using an RNN (recurrent neural network), conditioned on the image and all previously predicted words, to produce sentences. Their system is trained in an end-to-end fashion, such that any detail and context can be revealed by the hidden structure.

A general advantage of multimodal embeddings is that they can be used in a number of applications, e.g., for image-sentence retrieval tasks by using ranking algorithms or for text generation by training a network above the embedding space. Ngiam et al. [16] show that multimodal embeddings learned via autoencoders can even enhance results on tasks that do not obviously incorporate more than one modality.

Some approaches also consider the generation of more realistic image captions, thus captions that do not state what is visually obvious, and aim to build a bridge that connects an image with its context. Ramisa et al. [18] report on various tasks on collected news articles including caption generation. Also Feng and Lapata [6] suggest an approach to generate *context-aware* captions on another news corpora. However, current captioning results in this field are poor, which may come from the loose relation between an article text and its image. Beside caption generation, they also propose a method to extract a sentence from the article text that can serve as a legitimate image caption. But in general it is doubtable that co-occurring text provides appropriate captions.

[3]Source: https://en.wikinews.org/wiki/Turing_test_beaten_by_Russian_chatterbot (Accessed: 2/3/17)
[4]Source: https://simple.wikipedia.org/wiki/Air (Accessed: 2/3/17)
[5]It is assumed that image and text are jointly placed *on purpose*.

There is a variety of other applications involving or leveraging multiple modalities (e.g., question answering [22]). For instance, Izadinia et al. [10] show that pure NLP tasks, i.e., paraphrase detection, can benefit from learning semantic correspondences of visual similar scenes.

While a larger number of proposals is exploiting more than one modality, only few works concentrate on a closer investigation of the relation between co-occurring image-text pairs and how to utilize this relation. Yanai and Barnard [25] are trying to estimate the uncertainty of how an image region will be affected by a concept using an entropy measure. Here, they directly want to estimate the *visualness* of adjectives. For instance, the word *dark* is considered to be more visual than the word *religious*, simply because there is less variability in how an image region can be modified by an associated concept such as *dark* compared to a concept, that does not as easily reveal its influence (e.g., a *religious* image region might still depict anything from churches to ancient vases).

There are also some attempts to model semantic correlation between images and texts [23, 26, 27]. Xue et al. [23] propose an approach to estimate semantic correlation by aligning the semantics of visual and textual blobs (local image regions and words) . In order to assign blobs to a document of another modality, they have to make the assumption that co-occurring image-text pairs do express the same semantics. This assumption allows them to transfer a distribution of hidden topics learned among entities from one modality to another one, such that, e.g., visual blobs can be assigned to a textual document. In this case, the assumption is true since they are utilizing an image tagging dataset, i.e., each image is tagged with words that have a high semantic relevance with respect to that image.

3 IMAGE-TEXT RELATIONS

As the analysis of related work reveals, different levels of semantic image-text interrelations have not been investigated yet. In this section, we provide an analysis of important aspects of image-text interrelations and derive two measures to appropriately capture their characteristics. In particular, we are interested in the question in which way visual and textual data complement one another. Humans are involved at both sides of the communication channel: Humans intentionally add visual information to text (or vice versa) in order to supplement additional (normally complementary) information, and humans perceive and interpret such kind of bi-modal information. Of course, the intended effect is not always achieved and depends on many aspects (e.g., knowledge of creator and viewer etc.). Human knowledge about textual facts and depicted visual content plays a vital role in this process of communication. Before we introduce two measures to describe image-text interrelations, we discuss some examples of image-text pairs that share different kinds of information.

3.1 Examples for Image-Text Relations

The understanding of the image-text interrelations requires an analysis of 1.) what can be expressed by either of those modalities and 2.) how humans perceive and evaluate their co-occurrence.

Figure 2 shows the rare case, that a text and an image actually have the same information content. If each entity of one modality

A gray circle with a diameter of 1cm.

Figure 2: Example of a text and an image that convey the same information.

would have a corresponding entity in the other modality that exactly contains the same information, then one might claim either one of the modalities is obsolete. However, it is easier for humans to perceive attributes such as shape and color from an image, whereas the exact size is easier to read from a text. In fact, human capabilities of judging absolute measures of a visual object's length, size, area without any additional supporting information is rather limited. Hence, a certain modality can make the same information more difficult (or easier) to perceive. This leads to a natural usage where the depicted information shall complement one another such that each modality depicts partially unique information, that is easier to read from compared to other modalities.

Moreover, some kind of information cannot be encoded in one modality as succinctly and precisely as in another one. For instance, the sentence

Ada Lovelace was born in 1815.

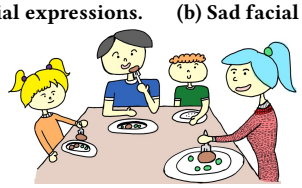
has no proper representation in an image without using text. An image that aims to convey the same information would necessarily express a lot of redundant additional information, such as events unique to the year 1815. The same applies for the reverse direction. There is neither a text that precisely describes the shape of a certain maple leaf nor its texture. Both observations lead to the conclusion that each modality plays an essential role to convey certain kinds of information by either addressing strengths or avoiding weaknesses of human visual perception.

Interrelations of images and texts can be understood as the alignment of concepts. Again, we consider an example:

A family of four is sitting at a table having a warm meal.
They are all talking vividly about their day.

The sentence equally fits to the images in Figure 3a and 3c, respectively. But when relating the sentence to Figure 3b, where the family is expressing a sad mood, this is intuitively perceived as a contradiction (or in other words: a negative correlation). The contradiction of concepts is that a sad mood normally is not aligned with a vivid conversation as we know from our own experience. However, the reason for human intuition when judging the interrelation of such co-occurrences is often not obvious nor easily expressible. More precisely, the alignments that define the interrelation are hidden. To some extent, this is similar to the problem of paraphrase detection, where humans are easily able to judge whether sentences express the same meaning but struggle to deliver a sensible and consistent reasoning for that claim in terms of syntactic and semantic justifications. While the annotation in the case of paraphrase detection is clear, it is not obvious how to quantify or rate the relation of image-text pairs as they usually do not represent the same meaning as outlined above. Moreover, they complement one another such that a good annotation would take into account the rationale or purpose of their co-occurrence.

(a) Happy facial expressions. (b) Sad facial expressions.



(c) Color of plates has changed.

Figure 3: The images (b) and (c) only differ in a certain concept from the image (a). The effect of the modification depends on the textual context.

3.2 Measures for Image-Text Relations

Three major goals are associated with the proposed measures for describing image-text relations. First, the measures should be easily explainable to humans, in particular for annotation and retrieval processes. Second, the descriptions, i.e., labels, for a given image-text pair should be easily inferable and disagreements between annotators should be avoided as far as possible. Third, the descriptions should be expressive, i.e., revealing distinguishable and high-quality relations. This goal enlarges the pool of possible future applications.

Judging several distinct aspects of complex inter-modal relations is easier than estimating a single score or classifying them into certain categories. This claim assumes that a one-dimensional categorization that meets the previously stated goals is not easily inferable. Therefore, we suggest to judge inter-modal relations based on two measurements, namely *Cross-modal Mutual Information* and *Semantic Correlation*.

Cross-modal Mutual Information (CMI) focuses solely on the mutual presence of concepts. Note, that the outlined measurement of *Cross-modal Mutual Information* is not equivalent to the term of mutual information in information theory. In order to better phrase its purpose, we take over the idea of *visualness* of concepts introduced in [2, 25] as explained in Section 2. Image captioning samples are a prominent example for high CMI for two reasons. First, the text exclusively focuses on the image content[6], thus only a few concepts are depicted solely in the image. Second, the shared concepts are highly visual, meaning that they have clearly defined representations in both modalities. On the other hand, image-text pairs occurring in news articles are usually characterized by a rather loose intersection in terms of information content. For instance, an article about austerity politics associated with an image depicting a piggy bank shares only one concept *"saving money"*. Hence, the amount of shared concepts is low, but also the visualness of the concept is low. A concept that has low visualness (e.g., "spring", "saving money", "religious") tolerates an immense variety among possible visual representations. The detection of concepts with low visualness is particularly difficult as it requires extensive background knowledge. However, visualness alone is not a sufficient measure to judge the salience of a concept. Even highly visual concepts might represent negligible details depending on the context. Hence, the annotator has to estimate the amount of shared concepts as well as their influence and importance.

Cross-modal Mutual Information solely does not sufficiently describe inter-modal relations. Irrespective of the amount of shared concepts, the appearance of only one pair of contradicting concepts might lead to an unfitting or disturbing image-text relation. Therefore, we propose a second measure called ***Semantic Correlation***, which aims to reveal how much meaning a text and an image share. This measure aims to mimic human intuition with respect to the sophisticated ability to detect matching pairs by considering context and regardless of the amount of shared information. A negative score shall indicate that the co-occurrence of an image and a text disturbs the comprehension of the depicted information, whereas a positive score eases the transfer of knowledge. The measure can be illustrated as follows. If two entities do not have any concepts in common, they are considered as unrelated (no correlation). If concepts appear that contradict one another, the correlation shall be estimated as negative. Depending on the relevance of the contradicting concepts, the negative correlation might be low or high. For instance, a color might be wrongly stated in the text. If this incorrectly referred object does only play a minor/negligible role compared to the overall content, the comprehension task is only insignificantly perturbed and even positive correlation can be assigned. We suggest to use an interval of [0,1] for CMI and [-1, 1] for SC, respectively, and refer to the description of our annotation process in Section 4.4.

4 ESTIMATING IMAGE-TEXT RELATIONS

In this section, we describe in detail the proposed deep learning framework for automatically estimating image-text interrelations. The framework consists of two main components, an autoencoder and a classifier.

The overall goal of this work is to develop a system that mimics human intuition when judging the interrelation of co-occurring images and texts. Therefore, it is essential that the system is able to comprehend individual modalities and to correctly evaluate their coexistence. Our main incentive is that humans use and comprehend several modalities to convey information that complements one another. This insight has been already stated by others (e.g., [1]), but it has been weakly addressed by related work as outlined in Section 2. There are two reasons for that. First, it is very difficult to model human intuition that includes visual perception as well as complex cognitive processes. Second, immense computational power is necessary to process a sufficient amount of data to learn appropriate models to achieve at least a basic understanding of the world, which is necessary when considering multimodal documents from unconstrained domains.

The human learning process is twofold: supervised and unsupervised. We are observing the world and draw our own conclusions, but we get also directed and corrected by the people surrounding us. For instance, if we observe *elephants*, we are capable of extracting

[6]This would also allow us to view the CMI relation of captioning samples as an inclusion, since the text does not express concepts that are not covered by the image.

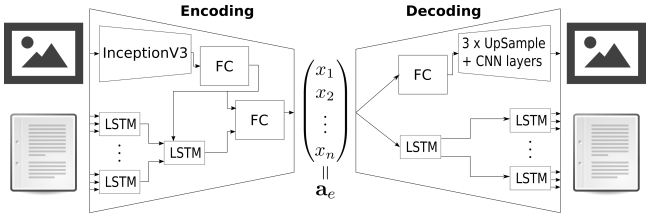
prominent features (shape, trunk, skin color, etc.) and generalize all elephants into a single concept without supervision. However, someone has to tell us that these mammals are called *"elephants"*. Hence, the overwhelming majority – but also the apparently more complicated – part of the learning task is done fully unsupervised. Nonetheless, this insight is encouraging since it allows us to train a complex system with just a small fraction of supervised training data or intervention, respectively. Still, annotated training data is necessary to direct the learning process such that the semantic outcome is aligned with our understanding of the world.

4.1 The Autoencoder Network Structure

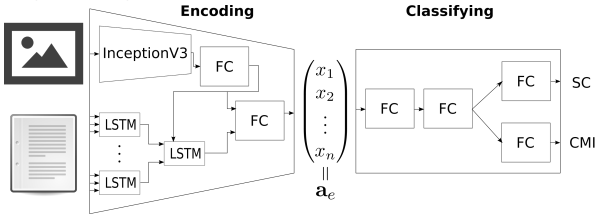
As it has been highlighted during the description of our measurements (CMI and SC), extensive knowledge about the world is required to quantify the co-occurrence of images and texts. More precisely, concepts have to be generalized within and across modalities. For instance, synonyms and paraphrases in sentences and texts, respectively, have to be identified as well as objects and actions in images. The generalized concepts within modalities have to be mapped to a multimodal representation.

A supervised scheme would require an infeasible amount of annotated data – representing as many constellations of image-text relations as possible – to accomplish this goal. Therefore, we propose to learn this ability via an unsupervised learning scheme. One obvious proper realization of this learning scheme would be through GANs (generative adversarial networks) [9], as they are uniquely capable of learning the semantics rather than the syntax of the input space. For instance, Radford et al. [17] have shown that GANs allow vector arithmetics on images similar to word embeddings. However, since we intend to directly learn feature vectors, we decided to use an autoencoder architecture instead. The encoder network compresses the input to a low-dimensional representation that contains less redundant information. Subsequently, a decoder network decompresses this intermediate representation back to the original input encoding. The intermediate representation can be considered as a feature vector that describes the complete input in a vector space of lower dimension. To achieve this, the encoder has to generalize concepts (e.g. objects, shapes, or poses in images) that are available in the input data.

As a point of departure, we use the implementation of Vinyals et al. [20] to build our own model. Figure 4a depicts the autoencoder. Image-text pairs are mapped to an intermediate feature representation, called article embedding $\mathbf{a}_e$. The decoding network tries to restore the initial input from this embedding $\mathbf{a}_e$. To encode images, we leverage *InceptionV3* [19] followed by a fully-connected layer (FC) to generate a final image embedding. All FC and LSTM (long short-term memory) layers in our networks use dropout for regularization. Texts are embedded in a hierarchical LSTM network that considers a sentence as a sequence of words and a text as a sequence of sentences. To generate a proper initialization of word embeddings, a *Word2Vec* model [15] has been trained prior to the autoencoder training. The network can basically adjust to dynamically sized sentences and text, respectively (although sentences are constrained by a maximum length). The hierarchical structure enables a more natural way of text processing since it allows us to consider sentences as self-contained entities. Furthermore, it is



(a) The autoencoder encodes an image-text pair into a compact representation. Subsequently, a decoder network tries to restore the original image-text pair.



(b) The classifier uses the encoding architecture from the autoencoder to map an input pair to a multimodal embedding. A multi-class classifier on top of that embedding quantifies the image-text relation.

Figure 4: The simplified architectures of the autoencoder (a) and classifier (b).

doubtable that a single LSTM layer can process a text as a plain sequence of tokens, as it is more difficult to maintain long-term dependencies if the input sequence becomes too long. The last output that is generated by an LSTM layer for a given input sequence is considered as a sentence or text embedding, respectively. To generate the text embedding, the first input to the LSTM layer is the image embedding in order to emulate a natural article processing (reading the text under consideration of the enclosed image). After the whole text has been processed, the image is reconsidered due to a further FC layer, that produces the final article embedding $\mathbf{a}_e$.

Ideally, $\mathbf{a}_e$ can be decompressed by the decoder network without loss of information compared to the original input. The basic architecture of the decoder network is depicted on the right side of Figure 4a. The decoder is split into two networks that receive $\mathbf{a}_e$ as input.

The upper part of the decoding network represents the image decoding. The image decoder computes a thumbnail via a fully-connected layer to extract visual information from $\mathbf{a}_e$. This thumbnail is then gradually up-sampled and refined through a series of CNN layers until the size of the input image is reached. More precisely, the network consists of three up-sample layers, each followed by a convolutional layer. The up-sample layers use nearest neighbor interpolation to increase the input size. The consecutive CNN layers are using 32, 8, and 3 feature maps, respectively. A squared-error loss evaluates the prediction compared to the input image.

The lower part of the decoding network depicts the text decoder. The text decoding architecture is reverse to the text encoding architecture. An LSTM layer generates a sequence of predicted sentence

embeddings. Therefore, at each time step it takes the article embedding $\mathbf{a}_e$ as input in addition to its previous state. Analogously, predicted sentence embeddings are decoded into tokens. The text decoding network does not allow dynamically sized predictions. The same restriction applies for the number of tokens in predicted sentences.

To estimate the quality of the predicted token embeddings, they have to be retranslated to words of the vocabulary. This is done by computing the cosine similarity between the predicted word embedding and all embeddings of words in the vocabulary. A softmax layer followed by a cross-entropy loss is used to compare the token predictions with the input text.

4.2 The Classifier Network Structure

The classifier combines the already accomplished achievements. Therefore, annotated samples are mapped to a feature representation via an encoder network, which has been learned by the autoencoder. Subsequently, a classifier network (left-hand side of Figure 4b) tries to infer CMI and SC labels for the sample. Recall, that the feature representation ideally contains all the information comprised by the input entities, as the features were trained with the ability to restore this input from them. Hence, the feature representation can be viewed as machine-readable representation of the sample, that hopefully allows an elementary concept matching even for non-visual concepts.

The encoding network is initialized via a pre-trained autoencoder model. In the best case, the encoder network does not require further weight modifications. Remember, that we justify a small annotated dataset with the unsupervised learning of strong feature representations. So, the supervised process is only needed to learn the relatively small classifier network on top of a complex encoder network. The encoded sample $\mathbf{a}_e$ is processed by a series of two fully-connected layers, that shall reveal the hidden alignments between textual and visual concepts. Subsequently, two separate fully-connected units are computing the final MI and SC prediction, respectively. As an alternative, we will also evaluate a SVM (Support Vector Machine) implementation [4] for classification.

4.3 Datasets

To meet the claims with respect to comprehending complex intermodal relations and to appropriately train the autoencoder and classifier, a diverse training database is needed. This database should sufficiently encode knowledge about the world, represent natural co-occurrences and enable the understanding of the semantics of images and text. For this purpose, we have leveraged three different datasets.

The first dataset shall enable the system to learn a translation of salient information from one modality to another. Therefore, the image captioning dataset MS COCO [3] is used, since such a dataset uniquely represents alignments of highly visual concepts between both modalities. The BBC news article set from [5] is used as an example corpus of particularly complicated image-text relations, since their content is typically loosely correlated and the meaning of their co-occurrence usually hard to infer. In most cases, there are neither direct references in the text to the image content nor is their semantic correlation easily inferable. Even humans do often need the provided caption to understand why the image fits to its article.

In addition, a dataset of encyclopedia articles is included in order to incorporate knowledge about the world. An online-encyclopedia such as Wikipedia is a powerful knowledge base that at the same time is structured and sufficiently trustworthy. Wikipedia contains general knowledge about the world and its entities, but also specialized knowledge about individuals, historic as well as recent events, or even proprietary products. However, many articles are not or at least difficult to understand for someone who is outside the subject area. Therefore, we have decided to use Simple English Wikipedia[7] (SimpleWiki) instead of the more extensive but also more complex English Wikipedia. SimpleWiki is the same as the normal Wikipedia, except that it aims to convey complex matters with simple textual descriptions. Such an encyclopedia dataset is necessary, since the understanding of relations between different modalities often requires background knowledge. We have created an encyclopedia dataset, that we call *SimpleWiki* dataset, by downloading articles from SimpleWiki and generating image-text pairs. We allow all occurring image types, such as photos, charts, maps, and drawings. An image is either matched with the text of its enclosed section or with the text of the full article in case when the image is associated with the summary. Currently, our SimpleWiki dataset consists of 2,999 image-text pairs.

4.4 The Annotation Process

Annotations have been gathered for subsets of all three datasets described in section 4.3. Although we basically allow real-valued data for both measures, we have simplified the annotation process and used only five different levels for each. The distribution of labels is shown in Table 1 and 2). In addition to judging *Cross-modal Mutual Information* and *Semantic Correlation*, text snippets were marked that can be considered as specifically relevant given the current image and to select the image type of the depicted image (e.g., photograph, map, chart, etc.).

In total, **761** annotations have been generated for the BBC News (205) and SimpleWiki (556) dataset by one of the authors. We have defined detailed label descriptions and examples in order to precisely judge the intermodal relation in the desired and a reproducible manner. The exact distribution of image types among those datasets shows that our overall dataset is still biased towards photos, because 71% of the SimpleWiki and 97% of the news samples are photos. In future work, this imbalance may be addressed by incorporating scientific articles into the dataset.

Since the MS COCO dataset has a homogeneous image-text-relation type by means of our measures, we have assigned high CMI and SC values to the 100 samples taken from this dataset. This step has been undertaken to reduce a strong label imbalance among CMI labels, since in natural image-text co-occurrences the text normally does not state obvious visual facts. To prevent the system from overfitting by learning the length of the text (an image caption is always a single sentence), we have concatenated a random subset of all 5 provided reference captions to generate an image-text pair.

[7] https://simple.wikipedia.org

Label	**0**	**0.25**	**0.5**	**0.75**	**1.0**
Meaning	$T \cap I = \emptyset$	$T \cap I \neq \emptyset$			
# Samples	44	157	466	52	107

Table 1: Distribution of CMI labels in the newly annotated dataset.

Label	**-1.0**	**-0.5**	**0.0**	**0.5**	**1.0**
# Samples	7	31	109	138	541

Table 2: Distribution of SC labels in the newly annotated dataset.

The final distributions of *Cross-modal Mutual Information* and *Semantic Correlation* labels are shown in Table 1 and Table 2, respectively. As stated in the previous section, most image-text pairs share concepts of both types, abstract (e.g., spring) or highly visual. This is the reason why there are more sublevels between 0 and 1 for CMI. In the annotation process, the amount of shared concepts has to be rated from label 0 (no intersection) to label 1 (large intersection) based on definitions for each case. Some samples have been marked as invalid, because the automatic retrieval of SimpleWiki samples has led to meaningless text extractions in rare cases. Altogether, 826 pairs have been sampled to generate a dataset for the classification scenario.

However, our initial claim, that the semantic correlation of co-occurring image-text pairs is not necessarily high, has been verified. News articles in our dataset have an average *Semantic Correlation* of 0.15, whereas SimpleWiki articles have an average SC of 0.88.

5 EXPERIMENTAL RESULTS

In this section, experimental results are presented for the proposed approach relying on a deep learning architecture to judge image-text relations. All experiments have been conducted using the system explained in Section 4, the dataset described in Section 4.3, and the annotated subset explained in Section 4.4, respectively.

5.1 Experimental Setup

Autoencoder (AE) and classifier (CL) are using stochastic gradient descent (SGD) with mini-batches and an initial learning rate of 0.1. The learning rate is halved every time a complete sweep through the training set has been accomplished. A mini batch consists of 16 image-text pairs. Note, that all samples within a batch are padded to have the same size as the largest sample. To further reduce this maximum size, texts have been truncated during preprocessing. We have found out that a maximum text size of 50 sentences and a maximum sentence length of 40 tokens yielded a manageable memory utilization per batch. This restriction does not severely distort the sample texts since only a few samples are affected by this measure.

In order to be included in the vocabulary, a word has to appear at least 10 times in the AE training set. Furthermore, a dictionary has been used to translate words from British English to American English for all samples taken from the *BBC News Database*. In this way, the vocabulary could be reduced from its original size of 59, 349 tokens to a final size of 12, 591.

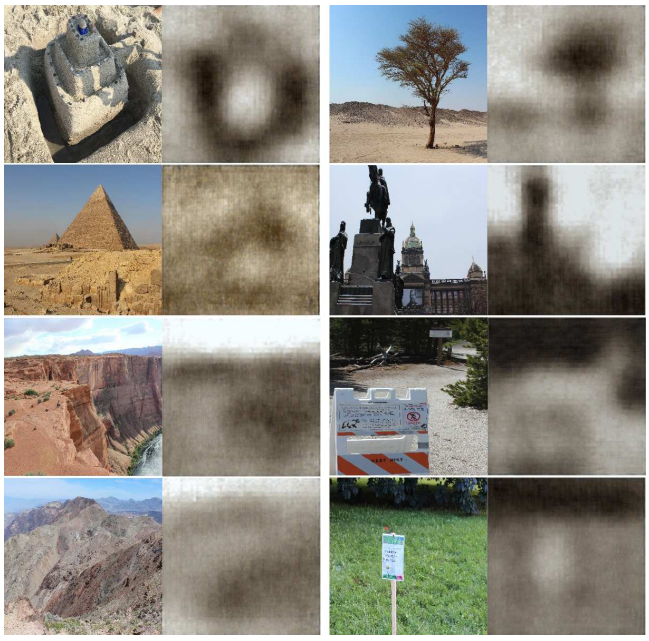


Figure 5: Example input-output image pairs of the trained autoencoder. These are randomly chosen unseen samples, i.e., they have not been seen during training.

The complete AE dataset is decomposed into three parts. 202, 654 samples have been generated from the MS COCO validation set (all image-caption pairs). In addition, all 3, 361 image-text pairs from the BBC News Corpora and 2, 999 image-text pairs from the SimpleWiki dataset have been included. From this randomly shuffled corpus, samples have been selected to generate a disjoint split of 190, 202 training and 6, 270 validation samples[8]. The image encoding network has been initialized with weights of a pre-trained *InceptionV3* model. Initial word embedding estimates have been taken from a *Word2Vec* implementation that was trained among the whole text contained in the dataset.

As outlined in Section 4.4, the CL dataset consists of 826 samples. The dataset has been divided in a training set consisting of 734 samples and a test set consisting of 92 samples. The CL encoding network has been initialized with the weights learned during AE training. Both systems use 300-dimensional word embeddings, 600-dimensional sentence and image embeddings, as well as 2400-dimensional article embeddings. Input images are scaled to size 300×300.

5.2 Performance of the Autoencoder

The capabilities of the AE are depicted in Figure 5. To make a qualitative statement about its performance, we measure the perplexity. During training, the image perplexity has decreased by 16.6% and the text perplexity by 5.5%, respectively.

As it can be seen, the AE is capable to store the global image contours in the extremely dense intermediate image embedding.

[8]The remaining samples are allocated for future usage.

Experiment	Accuracy CMI	Accuracy SC
CL	**0.6953**	**0.7344**
$\mathbb{E}_{\text{CL}}^{\text{no } AE}$	0.5625	0.6562
$\mathbb{E}_{\text{SVM}}^{AE}$	0.6875	0.7125
$\mathbb{E}_{\text{RAND}}^{MF}$	0.5642	0.6550

Table 3: The overall accuracy of predicting the correct CMI and SC labels that has been achieved in our experiments. The first row contains the results of the trained CL model.

Small architectural improvements may be sufficient to represent the salient semantics, such that the decoded image can be interpreted without the need of knowing the original input.

However, the text encoding has not been as successful yet. This may be due to the careful engineering of the utilized *InceptionV3* model which has no counterpart in the text decoding network. Yet, it can be assumed that the AE architecture is suited for feature learning and especially for conceptualization.

5.3 Performance of the Classifier

The encoder network of the CL has been initialized with the AE weights from the previous section. The article embeddings generated with the AE encoder do not fully contain the salient semantics yet. Therefore, the encoder network for article embeddings and the classifying network (Figure 4b) have been trained jointly in the supervised learning process. To minimize the risk of overfitting in this setting, we omitted one intermediate FC layer of the classifying network in Figure 4b. Furthermore, we stated the prediction of CMI and SC as multiclass problems using a cross-entropy loss.

In addition, the following systems have been setup as reference baselines for comparison:

- $\mathbb{E}_{\text{CL}}^{\text{no } AE}$: The trained classifier CL is used, but with randomly initialized weights in the encoder network, i.e., pre-trained AE is not used.
- $\mathbb{E}_{\text{SVM}}^{AE}$: A multiclass SVM [4], trained with the feature vectors of article embeddings $\mathbf{a}_e$ that have been generated by the trained AE model from Section 5.2[9].
- $\mathbb{E}_{\text{RAND}}^{MF}$: A random baseline, i.e., a "classifier" that simply outputs the most frequent label.

The accuracy achieved in all the previously described experiments is depicted in Table 3. The experimental results show that the deep learning architecture is basically able to predict image-text relations by means of CMI and SC. In both cases, the deep learning system outperforms the SVM approach. The results also reveal that the proposed pipeline consisting of an unsupervised pre-training and supervised refinement is necessary. Without the initialization of pre-trained weights, the classifier does not even outperform the random baseline.

6 CONCLUSIONS

In this paper, we have presented a novel approach to estimate the relations of co-occurring image-text pairs. Based on an analysis by which means interrelations can differ, we have derived two measures to describe image-text relations: Cross-modal Mutual Information and Semantic Correlation. Furthermore, we have proposed a deep learning architecture that consists of both an unsupervised as well as a supervised learning component. The purpose of the unsupervised autoencoder is to achieve a compact representation of multimodal image-text relations while at the same time minimizing the supervision efforts, i.e., reducing the number of required training samples. A deep neural classifier was trained using the autoencoder representation. In addition, we constructed several baseline systems to highlight the strengths of the designed system. The baseline systems have been consistently outperformed by the proposed deep learning system. Moreover, we highlighted the necessity of the full learning pipeline, consisting of unsupervised concept clustering and supervised concept-relation learning. Hence, the feasibility of the proposed deep learning system has been demonstrated for the challenging task of estimating image-text relations.

In future work, we are planning to improve the intermediate autoencoder representation by using a more sophisticated network structure. An expressive article embedding may enable an alternative fully unsupervised approach that involves an estimate of the pointwise mutual information of two entities. Since probability estimates drawn from the initial modality representations are presumably not expressive enough, they can be computed from the feature distribution in article embeddings. Hopefully, this approach will resolve currently existing shortcomings due to insufficient size of annotated training data. Finally, we will improve the annotation process by employing a group of annotators and investigating in detail the level of subjective judgments by means of inter-coder agreement.

[9] A suitable value for *weight-decay* has been found via grid search.

REFERENCES

[1] Kobus Barnard, Pinar Duygulu, David Forsyth, Nando de Freitas, David Blei, and Michael Jordan. 2003. Matching Words and Pictures. *Journal of Machine Learning Research* 3, 2 (2003), 1107–1135.

[2] Kobus Barnard and Keiji Yanai. 2006. Mutual Information of Words and Pictures. *Information Theory and Applications* 2 (2006), 5 pages.

[3] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and Lawrence Zitnick. 2015. Microsoft COCO Captions: Data Collection and Evaluation Server. *arXiv* (2015), 7 pages. http://arxiv.org/abs/1504.00325

[4] Koby Crammer and Yoram Singer. 2002. On the Algorithmic Implementation of Multiclass Kernel-based Vector Machines. *Journal of Machine Learning Research* 2, 12 (2002), 265–292.

[5] Yansong Feng and Mirella Lapata. 2008. Automatic Image Annotation Using Auxiliary Text Information. *Proceedings of Association for Computational Linguistics* 8 (2008), 272–280.

[6] Yansong Feng and Mirella Lapata. 2013. Automatic Caption Generation for News Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 4 (2013), 797–812.

[7] Andrea Frome, Greg Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc' Aurelio Ranzato, and Tomas Mikolov. 2013. DeViSE: A Deep Visual-Semantic Embedding Model. *Proceedings of Neural Information Processing Systems* 26 (2013), 2121–2129.

[8] Yunchao Gong, Liwei Wang, Micah Hodosh, Julia Hockenmaier, and Svetlana Lazebnik. 2014. Improving Image-Sentence Embeddings Using Large Weakly Annotated Photo Collections. *Proceedings of European Conference on Computer Vision* 13 (2014), 529–545.

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. *Proceedings of Neural Information Processing Systems* 26 (2014), 2672–2680.

[10] Hamid Izadinia, Fereshteh Sadeghi, Santosh Divvala, Hannaneh Hajishirzi, Yejin Choi, and Ali Farhadi. 2015. Segment-Phrase Table for Semantic Segmentation, Visual Entailment and Paraphrasing. *Proceedings of the IEEE International*

Conference on Computer Vision (2015), 10–18.

[11] Andrej Karpathy, Armand Joulin, and Fei-Fei Li. 2014. Deep Fragment Embeddings for Bidirectional Image Sentence Mapping. *arXiv* (2014), 9 pages. http://arxiv.org/abs/1406.5679

[12] Andrej Karpathy and Fei-Fei Li. 2014. Deep Visual-Semantic Alignments for Generating Image Descriptions. *arXiv* (2014), 17 pages. http://arxiv.org/abs/1412.2306

[13] Wei Liu and Xiaoou Tang. 2005. Learning an Image-word Embedding for Image Auto-annotation on the Nonlinear Latent Space. *Proceedings of ACM International Conference on Multimedia* 13 (2005), 451–454.

[14] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan Yuille. 2014. Explain Images with Multimodal Recurrent Neural Networks. *arXiv* (2014), 9 pages. http://arxiv.org/abs/1410.1090

[15] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. *Proceedings of Neural Information Processing Systems* 26 (2013), 3111–3119.

[16] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Ng. 2011. Multimodal Deep Learning. *Proceedings of International Conference on Machine Learning* 28 (2011), 689–696.

[17] Alec Radford, Luke Metz, and Soumith Chintala. 2015. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *arXiv* (2015), 16 pages. https://arxiv.org/abs/1511.06434

[18] Arnau Ramisa, Fei Yan, Francesc Moreno-Noguer, and Krystian Mikolajczyk. 2016. Breakingnews: Article Annotation by Image and Text Processing. *arXiv* (2016), 21 pages. http://arxiv.org/abs/1603.07141

[19] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2015. Rethinking the Inception Architecture for Computer Vision. *arXiv* (2015), 10 pages. http://arxiv.org/abs/1512.00567

[20] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2014. Show and Tell: A Neural Image Caption Generator. *arXiv* (2014), 9 pages. http://arxiv.org/abs/1411.4555

[21] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2016. Show and Tell: Lessons Learned from the 2015 MSCOCO Image Captioning Challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, 4 (2016), 652–663.

[22] Qi Wu, Chunhua Shen, Lingqiao Liu, Anthony Dick, and Anton van den Hengel. 2016. What Value do Explicit High Level Concepts have in Vision to Language Problems? *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016), 203–212.

[23] Jiao Xue, Youtian Du, and Hanbing Shui. 2015. Semantic Correlation Mining between Images and Texts with Global Semantics and Local Mapping. *Proceedings of International Conference on Multimedia Modeling* 8936 (2015), 427–435.

[24] Fei Yan and Krystian Mikolajczyk. 2015. Deep Correlation for Matching Images and Text. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2015), 3441–3450.

[25] Keiji Yanai and Kobus Barnard. 2005. Image Region Entropy: A Measure of Visualness of Web Images Associated with one Concept. *Proceedings of the annual ACM international conference on Multimedia* 13 (2005), 419–422.

[26] Yi Zhang, Jeff Schneider, and Artur Dubrawski. 2008. Learning the Semantic Correlation: An Alternative Way to Gain from Unlabeled Text. *Proceedings of the International Conference on Neural Information Processing Systems* 21 (2008), 1945–1952.

[27] Yue-Ting Zhuang, Yi Yang, and Fei Wu. 2008. Mining Semantic Correlation of Heterogeneous Multimedia Data for Cross-Media Retrieval. *IEEE Transactions on Multimedia* 10, 2 (2008), 221–229.